NON-PARAMETRIC BOUNDS ON THE NEAREST NEIGHBOR CLASSIFICATION ACCURACY BASED ON THE HENZE-PENROSE METRIC

Sally Ghanem[⋆] *Erik Skau*[⋆] *Hamid Krim*[⋆] *Hamilton Scott Clouse*[†] *Wesam Sakla* [†]

[⋆] North Carolina State University, Raleigh, NC 27695, USA
[†] Air Force Research Lab, Wright-Patterson AFB, OH 45433 , USA

ABSTRACT

Analysis procedures for higher-dimensional data are generally computationally costly; thereby justifying the high research interest in the area. Entropy-based divergence measures have proven their effectiveness in many areas of computer vision and pattern recognition. However, the complexity of their implementation might be prohibitive in resource-limited applications, as they require estimates of probability densities which are very difficult to compute directly for high-dimensional data. In this paper, we investigate the usage of a non-parametric distribution-free metric, known as the Henze-Penrose test statistic, to estimate the divergence between different classes of vehicles. In this regard, we apply some common feature extraction techniques to further characterize the distributional separation relative to the original data. Moreover, we employ the Henze-Penrose metric to obtain bounds for the Nearest Neighbor (NN) classification accuracy. Simulation results demonstrate the effectiveness and the reliability of this metric in estimating the inter-class separability. In addition, the proposed bounds are exploited for selecting the least number of features that would retain sufficient discriminative information.

***Index Terms*—** Dimensionality reduction, classification, divergence measures, nearest neighbor graph, pattern recognition.

1. INTRODUCTION

Working with high-dimensional data spaces is generally very complex due to the well-known “curse of dimensionality”[1] which was first introduced by Richard E. Bellman when he was considering problems in dynamic optimization. Studying the nature of these information-rich environments has been a recurrent topic in computer vision, statistical pattern recognition, machine learning, speech processing, graph theory and signal/image processing. This continued interest can be attributed to the continuing growth in sensors’ capabilities.

Information-theoretic divergence measures have been used in many applications including but not limited to image registration [2, 3, 4], image segmentation and retrieval [5], image alignment [6], speech classification [7] as well as a variety of other problems. Examples of popular divergence measures include the Kullback-Liebler divergence [8] which is based on Shannon entropy measure and the so-called Renyi-divergence (alpha-divergence) measure [9] which is based on Renyi entropy. Other divergence measures include the Jensen-Shannon divergence [10], the Jensen-Renyi divergence [11], total variation k-dP divergence [12] and Bregman divergence [13].

Unfortunately, a common problem with most of these measures is, again, estimating the probability densities for the high-dimensional data given the complexity of computing the local estimate of the density. This problem has precipitated some restrictions for different applications. In 1979, Friedman and Rafsky [14] proposed the idea of using minimum spanning trees (MST) to extend the Wald-Wolfowitz test [15], which is also known as the two-sample test, for high dimensional data. We are going to introduce a new approach that relies on the Henze-Penrose [16] divergence measure, which is itself based on Friedman and Rafsky work, to estimate the number of features that would retain sufficient discriminative information. Using this criterion, the performance of the classification procedure is evaluated, in terms of accuracy, as a function of the number of dimensions. Bounds for the nearest neighbor [17] classification accuracy are derived in terms of the Henze-Penrose metric and exploited to validate our results.

The paper is organized as follows: Section 2 introduces the Friedman-Rafsky and Henze-Penrose test statistics and provides an overview of their properties. Section 3 introduces the bounds for the NN classification accuracy. Section 4 describes our dataset structure and the experimental setup. Section 5 presents the simulation results of the proposed bounds. Section 6 provides concluding remarks and a discussion of future work.

2. FRIEDMAN-RAFSKY AND HENZE-PENROSE TEST

Consider two classes w_0 and w_1 over the space X with samples of size m and n respectively from distribution $p(x)$ and $q(x)$, both defined on $\mathbb{R}^d$ where d is a positive integer number. According to the Wald-Wolfowitz test, the null hypothe-

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

sis H_o specifies that $p(x) = q(x)$ which means that both samples are drawn from the same underlying distribution. Our interest is in the case $H_1 : p(x) \neq q(x)$ where each sample belongs to a different class. The Wald-Wolfowitz test (for $d = 1$) begins by sorting the univariate observations $N = m+n$ in an ascending order with respect to their values. Each observation is then replaced by a label w_0 or w_1 depending upon the class to which it originally belonged. The number of runs, $R_{m,n}$, is the number of consecutive sequences of identical labels.

Friedman and Rafsky developed a graph-theoretic generalization for the univariate Wald-Wolfowitz statistic to compute the number of runs for higher dimensional data using MST. They proposed using the MST as a multi-variate extension for the two-sample test and they proved the appropriateness of this idea in [14]. Given an edge weighted graph consisting of the N pooled sample data points in $\mathbb{R}^d$ as nodes. The weight associated with each edge is a measure of dissimilarity between the nodes defining that edge, e.g. the Euclidean distance. The MST of this graph is thus the subgraph of minimum total distance that provides a path between every pair of nodes. The test statistic $R_{m,n}$ is now the number of connected components left after removing edges connecting different classes in the MST. $R_{m,n}$ provides a simple, yet effective non parametric measure of separation between the two samples by making use of the local characteristics of the distributions. Lower values of $R_{m,n}$ correspond to increased separation between the class distributions and vice versa.

Henze and Penrose [16] extended the work of Friedman and Rafsky by proving that as the number of vertices $m, n \longrightarrow \infty$, a function of the statistic $R_{m,n}$ asymptotically converges to a member of the f-divergence family as shown in Eq. (1). The Henze-Penrose divergence measure HP estimates the distributional overlap between two distributions $p(x)$ and $q(x)$ such that a and $b \in$ [0,1] where $a = \frac{m}{m+n}$ and $b = 1 - a$.

$$HP = 1 - \frac{R_{m,n}}{m+n} \longrightarrow \int \frac{a^2p^2(x) + b^2q^2(x)}{ap(x) + bq(x)} dx \qquad (1)$$

almost surely.

Thus, given a distance or proximity measure, HP can provide a measure of separation between classes of objects in the original sensed representation space. HP = 0.5 implies that the densities $p(x)$ and $q(x)$ are drawn from the same underlying distribution. As HP increases, the densities $p(x)$ and $q(x)$ are increasingly separated to the point where HP attains its maximum value at 1.

3. BOUNDS ON THE NEAREST NEIGHBOR GRAPH CLASSIFICATION ACCURACY

The Henze-Penrose metric can be used to provide bounds for the NN classification accuracy. One advantage for the proposed bounds is that they are estimable with no prior knowledge for the underlying distribution of the dataset.

Theorem. *Let labeled classes of m and n points form a complete graph of unique distances. Given the Henze-Penrose metric HP between the two classes and the number of connected components in the nearest neighbor graph C, the nearest neighbor classification accuracy, A_{NN}, is bounded above and below by:*

$$2 * HP + \frac{2}{m+n} - 1 \leq A_{NN} \leq HP + \frac{C}{m+n} \qquad (2)$$

Proof. Suppose two classes of m and n points form a complete graph of unique distances. Therefore, the MST is unique and the nearest neighbor graph (NNG) will be a subset of that MST.

We proceed with some realizations relating the number of misclassified points, E, the number of edges that exist in the MST but not in the NNG, $C - 1$, and the number of edges connecting different classes in the MST, $R_{m,n} - 1$. Each edge connecting different classes in the MST causes at most two classification errors and each edge connecting different classes in the MST either does not exist in the NNG or causes at least one classification error.

$$\frac{E}{2} \leq R_{m,n} - 1 \leq C - 1 + E \qquad (3)$$

We next prove the lower bound by writing :

$$\begin{aligned} A_{NN} &= 1 - \frac{E}{m+n} \\ &\geq 1 - \frac{2(R_{m,n} - 1)}{m+n} \\ &\geq 1 - \frac{2((1 - HP)(m+n) - 1)}{m+n} \\ &\geq 2HP + \frac{2}{m+n} - 1 \end{aligned}$$

Lastly, we prove the upper bound by proceeding as :

$$\begin{aligned} A_{NN} &= 1 - \frac{E}{m+n} \\ &\leq 1 - \frac{R_{m,n} - C}{m+n} \\ &\leq 1 - \frac{(1 - HP)(m+n) - C}{m+n} \\ &\leq HP + \frac{C}{m+n} \end{aligned}$$

□

Practically, the lower bound is more useful since it does not require calculating the number of nearest neighbor components. It proves that as the HP metric tends to 1, so does the NN classification accuracy. Both the upper and lower bounds are tight and can not be improved upon without further assumptions.

4. EXPERIMENTAL VALIDATION

The synthetic imagery data-set we use consists of a variety of vehicles, some with high variability and others with high similarity. It is comprised of 7056 images for fourteen different vehicles; ten of them are civilian vehicles and the other four are military vehicles. Six of the civilian vehicles are sedans, two are sport utility vehicles (SUVs), one is a minivan and the other one is a pickup truck. Two of the military vehicles are treaded tanks and the other two are armored carriers with wheels. We divided our dataset into three classes as discussed in Table 1.

	Number of vehicles	Number of images
Class 1	6 sedans	3024
Class 2	2 SUVs and 2 trucks	2016
Class 3	4 military vehicles	2016

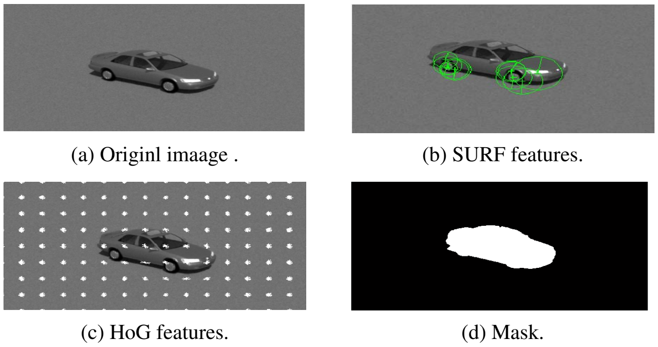
Table 1: The dataset description.

The images for each vehicle were collected at seven different elevations and seventy two different azimuth values such that each elevation level has 72 different viewpoints to represent the same vehicle. This resulted in 504 images for each vehicle. All the images were converted to gray-scale values and cropped to 76x76 pixels as pre-processing steps to increase the computational efficiency. The images were vectorized, $x_i = \text{vectorize}(Image_i)$, and the Euclidean distance, $d(x_i, x_j) = \|x_i - x_j\|_2$, was used as our proximity measure through all the experiments. The HP metric is computed for the original dataset between each pair of classes and the results are highlighted in Table 2.

We subsequently applied some common feature extraction techniques like Speeded Up Robust Features (SURF) [18] and Histogram of Gradients (HoG) [19], to assess their effectiveness in preserving separation between each pair of classes in our dataset. We also computed the gradient mask (or the silhouette) for the vehicles through contrasting the vehicle from the background. Changes in contrast can be detected by operators that calculate the gradient of the image. Furthermore, a threshold is applied to create a binary mask containing the segmented vehicle after filling the interior gaps inside the vehicle. The adopted features are shown in Figs. (1a-1d). Levels of separation between each pair of classes, represented by the HP metric, for the different feature spaces are depicted in Table 2.

5. EXPERIMENTAL RESULTS

We evaluated the inter-class separability versus the classification accuracy using two methods for classification : k-Nearest Neighbors (k-NN for k=1) [20] and Support Vector Machine (SVM) [21]. A representative one-third of the dataset was used to train each classifier and the testing was performed on the rest of the data. The accuracy of classification for each pair of classes was computed using the previously mentioned classifiers along with the inter-class HP metric. The results are shown in Figs. (2a-2d). As expected; the accuracy for both classifiers increases as the HP value increases. This is intuitive since as the separation between different classes increases, it becomes easier for the classifier to efficiently perform the discrimination task. Furthermore, the HP values and the NNG classification accuracy nicely track, which agrees with the bounds derived in Section 3.



(a) Originl imaage .

(b) SURF features.

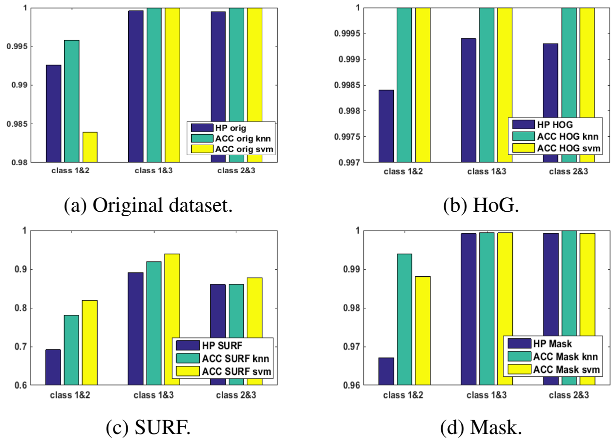
(c) HoG features.

(d) Mask.

Fig. 1: The different adopted features.

	Orig. data	SURF	HoG	Mask
HP_{12}	0.9926	0.6921	0.9984	0.9671
HP_{13}	0.9996	0.8915	0.9994	0.9992
HP_{23}	0.9995	0.8606	0.9993	0.9993

Table 2: The Henze-Penrose metric values for the different feature spaces.



(a) Original dataset.

(b) HoG.

(c) SURF.

(d) Mask.

Fig. 2: The Henze-Penrose metric versus the classification accuracy.

Looking at Fig. (2), we observe that the HoG descriptor, which gives special attention to geometric features, outperformed the other feature extraction techniques. The mask (or the silhouette), which only preserves the most general geometric features performed slightly worse than the original dataset. The SURF descriptor, which focuses more on regional features such as texture, was the worst performing feature extraction technique. This is reasonable because in our dataset geometric information is better suited than texture information for vehicle classification.

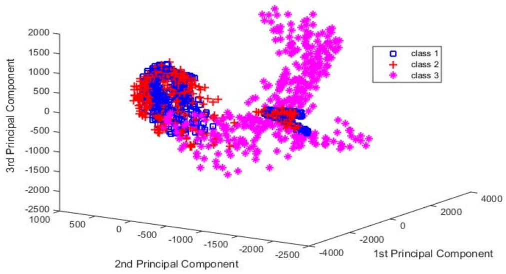
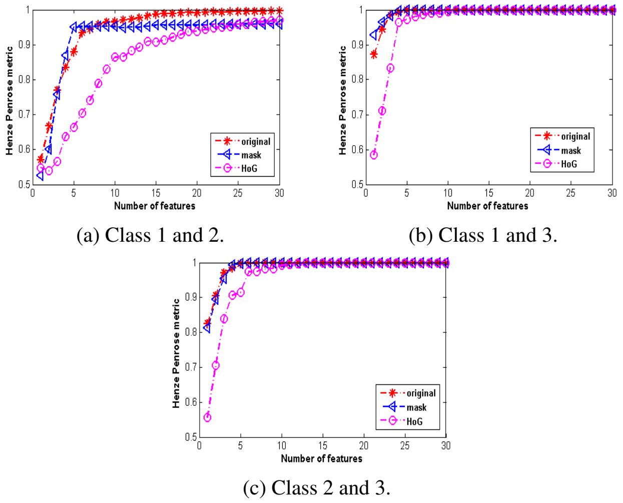


Fig. 3: The three-dimesional embedding for the original dataset using PCA.

We considered applying Principal Component Analysis (PCA) [22], a well-known linear dimension reduction technique, to the different feature spaces. The three dimensional embedding for the original dataset is shown in Fig. (3). We propose using the bounds derived in Section 3 to predict the number of features (or principle components) that would achieve a desired classification performance. Figs. (4a-4c) display the HP score for each pair of classes as a function of the number of principal components for the original dataset, HoG and mask. For a NNG accuracy of at least 0.9, we require a HP value greater than 0.9498 between classes 1-2 and 1-3, and a HP value greater than 0.94975 for classes 2-3. We selected the minimum number of features that would achieve our desired HP metric values. After choosing the number of principal components that would represent each image, we applied K-NNG classifier (k=1) on the reduced-dimension dataset to compute the actual classification accuracy. The results are shown in Table 3.

6. CONCLUSION

In this paper , we used a non-parametric metric to quantify the inherent separation between classes for a labeled set of high-dimensional synthetic vehicle imagery data. We also exploited various common feature descriptors to evaluate the performance of Henze-Penrose separation in the relevant feature spaces. The results demonstrated the robustness and the effectiveness of the Henze-Penrose metric through comparing the separation value versus the accuracy of classification for the raw data and in the various feature spaces, which emphasizes the role of this metric in evaluating the efficacy of feature extraction and dimensionality reduction mappings.



(a) Class 1 and 2.

(b) Class 1 and 3.

(c) Class 2 and 3.

Fig. 4: The Henze-Penrose metric between each pair of classes for different feature extraction techniques as a function of the dimensionality.

	No. of features	HP	Acc. low. bound	Actual accuracy.
Orig$_{12}$	8	0.9583	91.7%	97.46%
HoG$_{12}$	23	0.9502	90.01%	96.98%
Mask$_{12}$	6	0.9514	90.32%	97.38%
Orig$_{13}$	3	0.9845	96.93%	98.41%
HoG$_{13}$	4	0.9643	92.9%	96.87%
Mask$_{13}$	2	0.9653	93.1%	96.69%
Orig$_{23}$	3	0.971	94.25%	96.7%
HoG$_{23}$	6	0.9735	94.75%	98.02%
Mask$_{23}$	3	0.9531	90.67%	95.19%

Table 3: Results for NNG classification accuracy prediction.

Distribution-free metrics such as Henze-Penrose and Friedman-Rafsky provide an avenue to quantify the efficiency of a particular dataset with regards to its discrimination capability. It also affords the minimum number of reduced dimensions required to maintain discrimination that is comparable to the original dataset. Furthermore, using a simple yet effective non-parametric similarity measure would allow to incorporate prior knowledge regarding the performance of the NNG classifier.

Future work will focus on developing a mapping between the Henze-Penrose test and other classifiers. In addition, our goal is to extend the bounds on the k-NNG classification accuracy to include the cases where $k > 1$.

7. REFERENCES

[1] R. E. Bellman and H. A. Osborn, "Dynamic programming and the variation of green's functions.," 1957.

[2] A. B. Hamza and H. Krim, "Image registration and segmentation by maximizing the jensen-rényi divergence," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2003, pp. 147–163.

[3] B. Ma, A. Hero, J. Gorman, and O. Michel, "Image registration with minimum spanning tree algorithm," in *Image Processing, 2000. Proceedings. 2000 International Conference on*. IEEE, 2000, vol. 1, pp. 481–484.

[4] H. Neemuchwala, A. Hero, S. Zabuawala, and P. Carson, "Image registration methods in high-dimensional space," *International Journal of Imaging Systems and Technology*, vol. 16, no. 5, pp. 130–145, 2006.

[5] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 267–272.

[6] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 2, pp. 187–198, 1997.

[7] A. Wisler, V. Berisha, J. Liss, and A. Spanias, "Domain invariant speech features using a new divergence measure," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 77–82.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, pp. 79–86, 1951.

[9] A. Rényi et al., "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[10] J. Lin, "Divergence measures based on the shannon entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.

[11] Y. He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *Signal Processing, IEEE Transactions on*, vol. 51, no. 5, pp. 1211–1220, 2003.

[12] D. Stowell and M. D. Plumbley, "Fast multidimensional entropy estimation by k-d partitioning.," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 537–540, 2009.

[13] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.

[14] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *The Annals of Statistics*, pp. 697–717, 1979.

[15] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 147–162, 1940.

[16] N. Henze and M. D Penrose, "On the multivariate runs test," *Annals of statistics*, pp. 290–298, 1999.

[17] D. Eppstein, M. S. Paterson, and F. F. Yao, "On nearest-neighbor graphs," *Discrete & Computational Geometry*, vol. 17, no. 3, pp. 263–282, 1997.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[20] D. Coomans and D. L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules," *Analytica Chimica Acta*, vol. 136, pp. 15–27, 1982.

[21] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[22] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.